



TITLE:

新しいタイプの不完全データにもとづく統計的推測 (生命科学データの統計解析と数学モデル)

AUTHOR(S):

仮谷, 太一

CITATION:

仮谷, 太一. 新しいタイプの不完全データにもとづく統計的推測 (生命科学データの統計解析と数学モデル). 数理解析研究所講究録 1980, 384: 100-121

ISSUE DATE:

1980-05

URL:

<http://hdl.handle.net/2433/104853>

RIGHT:

新しいタイプの不完全データにもとづく統計的推測

川崎 匡大 仮谷 太一

0 生命科学実験とそのデータの特徴

生命科学データの統計解析においては, *random sampling* および標本データ概念に関して, 標準的なそれとの間に, 大きなギャップがあることを常に心掛けておく必要がある。理想的な実験母集団または調査母集団の構成は, 対象が複雑な履歴をもつ生体であるために, 本来不可能な場合が多く, 現実には得られる実験母集団または調査母集団は, 研究者のその問題に関する知識・経験の深浅, 技術の巧拙, 実験装置 および実験・調査組織などにより, 大きく左右されるからである。

例えば, いくつかの処理法の優劣を比較しようとする場合, できる限り均質と思われる対象個体を, 各処理に *random* に割りつけて実験を行ない, それぞれから得られるデータを, 各処理に対応する実験母集団からの無作為標本データとみなして統計解析にかける。この際,

個体の処理に対する反応と誘発する要因に関して、完全な情報を期待することは不可能であり、また分っている要因についても、それと必要に応じて管理または制御できないことが多く、実験・調査の技術や組織の問題ともかきんで、各処理に対応する実験母集団は、理想的な実験母集団とはかなり隔たっていることを覚悟しなくてはならない。

さらに標本個体の持つ特性値を直接計量しべたこと、中途脱落例の生ずること、生命科学データの特徴ということができるであろう。標本個体の持つ特性値と、それを表示する標本データとを区別し、中間に観測システムを設先して、そのシステムの最適化をもはかりながら、生命科学データの統計解析に取り組むことが肝要である。

① 臨床医学的実験または調査とその数学モデル

ある種の病気に対し、特定の医学的処理を行なっている、ある特性（再発、死亡など）が出現するまでの時間 T^* の分布を研究する問題を考えてみよう。

実験の目的ながらに計画を策定し、実際に研究が開始されるとき、実験対象となる患者は、その了解の下に、順次実験に参加することになる。実験開始直後から、実験終了の直前まで、つぎつぎに患者が実験に参加し、それぞれ何回かの診査を受けるというのが臨床医学的実験の普通の

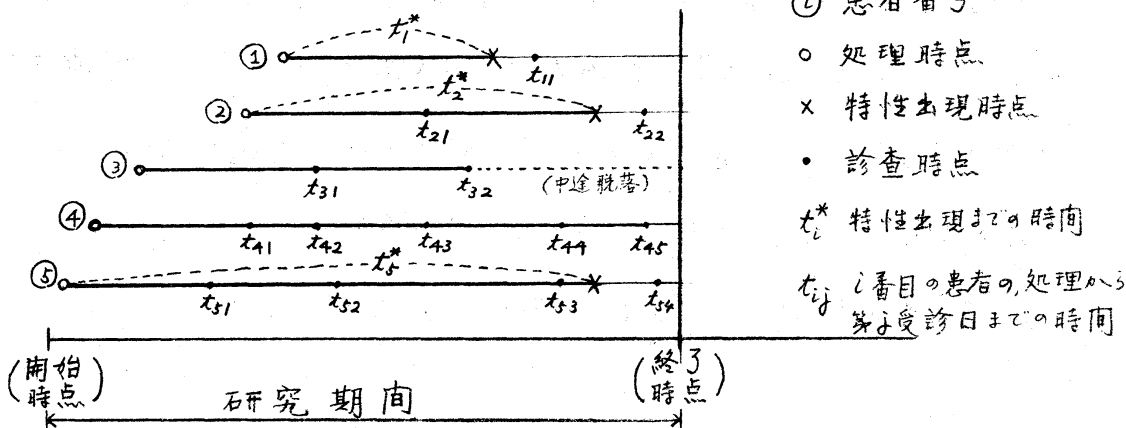
様式である。実験研究の期間中に、交通事故をはじめ、他の原因で死亡するとか、他の地域に転出するとか、あるいはまた、よく分らない理由で、実験から脱落するものがあることは避けられない。これらの脱落症例については、それぞれに特有なパターンが存在と否定することはできないであろうが、ここでは問題を単純化して、 T^* の分布に関しては、脱落症例も、追跡できた症例と全く同じ分布に従うものと仮定する。

なお、目指す特性が死亡などの場合には、かなり正確にその特性出現までの時間 T^* を観測することができると、標準的な統計解析を中心に、中途脱落した症例のもつ情報を追加し、一層信頼のおける結論を導くことに力点があかれる。より具体的にいえば、*exact* なデータの統計的推測に付加採用すべき、右に開いたセンサードデータの統計的取扱い方が、工夫の焦点である。こうしたデータの解析については、J. W. Boag [1] をはじめ、かなり多数のパラメトリックな研究報告、および数多くのノンパラメトリックな研究が知られている。しかしここでは、病気の再発、乳歯の萌出・歯蝕、初潮開始などのように、 T^* の値を、実際問題として、明確に観測することができず、個体ごとに定まるある時間区間 I に T^* が含まれることだけを知り得る場合について考察する。上記の時間区間 I を、以後

T^* の interval-censored data と呼ぶことにする。

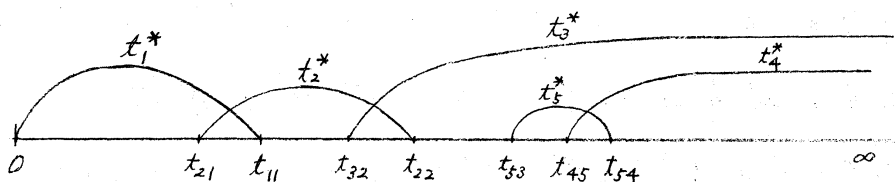
次の〈図1〉に数個の症例について、実験への参加、診査時点、特性出現時点などの概念図を示した。患者番号は、都合で実験への参加順とは逆になっている。

〈図1〉 臨床医学的実験の模式図



〈図1〉の5症例に対しては、 $t_1^* \in (0, t_{11}]$, $t_2^* \in (t_{21}, t_{22}]$, $t_3^* \in (t_{32}, \infty]$, $t_4^* \in (t_{45}, \infty]$, $t_5^* \in (t_{53}, t_{54}]$ が観測されるから、標本データとして、 $(0, t_{11}]$, $(t_{21}, t_{22}]$, $(t_{32}, \infty]$, $(t_{45}, \infty]$, $(t_{53}, t_{54}]$ なる5つの interval-censored data が得られる。これを念のため図示すると〈図2〉のようになる。

〈図2〉 Interval-censored Data



さて、上記の臨床医学的実験または調査にもとづく T^* については、中途脱落の症例も、研究終了のため特性出現まで追跡できなかった症例も、 T^* の確率分布に関する限り、特性出現時点観測可能症例と全く同一であるとの仮定の下で、次のような数学モデルを構築することができる。

まず、変数 X の変域を $(\alpha, \beta]$ とし、 x_1, \dots, x_n は $\alpha = x_0 < x_1 < \dots < x_n < x_{n+1} = \beta$ をみたすある数列とすると、関数 $I(X; x)$ 、 $x = (x_0, x_1, \dots, x_n, x_{n+1})$ を次のように定義しておく。

$X \in (x_{j-1}, x_j]$ のとき、そのときに限り

$$I(X; x) = (x_{j-1}, x_j] \quad \text{ただし } j = 1, 2, \dots, n+1. \quad (1)$$

確率変数 T^* は直接には観測することはできないが、その関数 $I(T^*; t)$ は常に観測可能である。ここに $t = (0, t_1, \dots, t_n, \infty)$ は T^* の値とは全く無関係に与えられる観測時点列 t_1, \dots, t_n に、 T^* の分布の下限 0, 上限 ∞ を付加して得られるベクトルで、 $0 < t_1 < \dots < t_n < \infty$ とする。

$F(t, \theta)$ を $(0, +\infty)$ 上のある分布関数とし、 Ω を $(0, a], (a, b], (b, \infty]$ ($0 < a < b < \infty$) なる形の区間からなるあるクラスとする。このとき T^* は $(0, \infty]$ と変域とする確率変数で、その確率分布は、

$$P_r(T^* \in (0, t]; \theta, \rho) = (1 - \rho) F(t, \theta), \quad 0 < t < \infty \quad |$$

$$\Pr(T^* = \infty; \theta, \rho) = \rho, \quad 0 \leq \rho < 1 \quad \} (2)$$

で与えられる。

なお、(2)における ρ は、臨床医学的には、ある医学的処理の結果、完全に治癒してしまう率（完治率）、また、齲蝕のような場合には、虫歯にならない率を表わしている。

このような数学モデルの下での、 T^* の分布に関する統計的推測の問題は、確率分布(2)に従う母集団からの、大きさ n の無作為標本 $(T_1^*, T_2^*, \dots, T_n^*)$ についての標本データが、*interval-censored data* $(I(T_1^*; t_1), I(T_2^*; t_2), \dots, I(T_n^*; t_n))$, $t_i = (0, t_{i1}, \dots, t_{ik_i}, \infty)$, $I(T_i^*; t_i) \in \Omega$ として与えられるような場合における、母数 θ, ρ に関する統計的推測の問題に帰着させることができる。

以後、標本変数 T^* と標本データ $I(T^*; t) \in \Omega$, $t = (0, t_1, \dots, t_k, \infty)$ に変換する観測システムを、*censoring observation system* と呼ぶことにする。ここに t_1, \dots, t_k は T^* の観測時点列を示し、 $0, \infty$ はそれぞれ T^* の分布の下限、上限である。

2 いくつかの実例

個体のある *quantal* な特性に関する医学的・疫学的調査では、ある特定の出発時点からその特性が出現するまでの時間 T^* は、直接には観測できず、個体ごとに定まるあ

る時間区間 I に, T^* が含まれることだけを知り得る場合が少なくない。そのような場合のいくつかをあげてみよう。

例1 初潮開始年令と, 小学校高学年の女子 または中学校の女生徒を対象に, ある日時に一斉調査を実施する場合。このとき それぞれについて, 初潮がすでに経験済みであるか否かだけを調査するとすれば, 得られるデータは, その時点における各自の年令に関し, 経験済みならば左に開いたセンサード・データ, まだならば右に開いたセンサード・データが得られる。

複数回 同じ対象に対して調査を繰り返すときは, 一般に *interval-censored data* が得られる。これまでの研究によると, 出生の年, 地域がほぼ同じである女子の初潮開始年令の分布は, 正規分布で近似できることが知られている.[4]。

例2 幼児の特定乳歯の萌出年令について, 数回集団診査を実施し, 萌出か否かと調査する場合。上下左右 20本の乳歯について, それぞれの萌出特点を観測記録することはきわめて困難と言わなければならない。さらに萌出ししかっている乳歯をどの時点で萌出とするかも技術的にかなり面倒な問題を含んでいる。実際問題としては, ある地域を定め, 同月出生児を対象に, 生後約1か年くらいから, 4か月おき または 6か月おきに実施されることが多いようである。なるべく医師は1人とし, 萌出の診査基準を厳格に守りながら実施しなくては, 信頼のおけるデータは得られない。こうして

得られる標本データは, *interval-censored data* で, 集団診査の回数が重なるほど脱落者が多くなり, また途中から新規に加入するもの, 数回の診査のうち何回かを受診しないものなど, いろいろである。しかし いずれにせよ, 全部について *interval-censored data* は入手することができ。乳歯の萌出年令分布については, 従来の研究から, 正規分布 または 対数正規分布で近似できることが知られている [6]。

例3 胸部がんの1種について, 外科的手術を行なったから, 再発するまでの時間と, その後の観察記録にもとづいて調査する場合。手術後どのように診察が行なわれるかは, その時期についても回数についても, 患者個々の事情によりまちまちである。また, 他の原因で死亡するとか, 他地域に転出するとか, あるいはまた他の病院に変わるとか, いろいろの要因により中途脱落症例が生じる。これらの要因別に患者を層別して考えれば, 再発までの時間に関して, 異なったパターンをいつかも知れない。さらに頻繁に診察に訪れる患者と, そうでない患者とでは, 再発までの時間に関して有意な差があるかも知れない。しかしこれらの詳細については, 簡単に説明することはできないので, われわれは そこに問題点のあることを指摘するに止め, 第1近似として, 全症例が同じ分布に従うものと仮定して 統計解析を行なわざるを得ないであろう。再発までの時間 T^* の分布については,

(2) のような確率分布で, $F(t, \theta)$ に対数正規分布を仮定する場合が多いようである。

例4 幼児の特定乳歯が萌出してから, 齲蝕発病までの時間を, 何回かにわたる継続調査で, 毎回萌出か否か, 齲蝕発病か否かと調査して, 研究する場合. この場合は, 幼児の出生から齲蝕発病までの時間ではなく, 特定乳歯が萌出してから, 齲蝕発病までの時間が問題であり, 出発点となる萌出時点も直接観測できず, *interval-censored data* としてしか与えられないという点で, 前の数例よりもさらに面倒である. 問題の時間について, 2次元的な接近も考えられるが, 齲蝕発病区間から, 萌出区間を引く算することにより, 区間の幅は広くなるけれども, それぞれに *interval-censored data* と算定することによって, 同じように統計解析にかけることができる. この萌出から齲蝕発病までの時間の分布については, (2) のような確率分布で, $F(t, \theta)$ を対数正規分布として近似することによって考えられる. なお, 下顎の前歯などは虫歯になりやすく, 個人により永久歯に代わるまで虫歯にならないものもあるので, (2) における ρ は正として取扱わなければならない。

[3] データの整理と最尤解

大きさ n の無作為標本 $(T_1^*, T_2^*, \dots, T_n^*)$ からある censoring observation system の下で得られた *interval-censored*

(9)

data $(I(T_i^*; t_i), \dots, I(T_m^*; t_m))^+$ の両端点をこみにして 小から大へと順に並べ, 必要ならば T^* の分布の下限 α , 上限 β を追加し, あらためて $\alpha \equiv t_0 < t_1 < \dots < t_m < t_{m+1} \equiv \beta$ とする. 各組 (i, j) ($0 \leq j < i \leq m+1$) に対し

C_{ij} = 左右の端点がそれぞれ x_j, x_i に等しい $I(T_i^*; t_i)$ の個数

を定義する. さらに次の条件:

$$(D_k) \quad \sum_{i=1}^{k-1} \sum_{j=0}^{i-1} C_{ij} + \sum_{j=k+1}^{m-1} \sum_{i=j+1}^m C_{ij} \neq 0 \quad (1 \leq k \leq m-1)$$

$$(D_0) \quad \sum_{i=2}^m \sum_{j=1}^{i-1} C_{ij} \neq 0$$

ただし, 上式で定義できない $\sum \sum C_{ij}$ は 0 とする

を定義すれば, $F(t, \theta)$ が正規分布, 対数正規分布などのとき, これらの条件を用いて, (θ, ρ) の最尤解が存在することを証明することができ, Newton-Raphson 法による最尤推定が可能であるが, ここにはその詳細は省略する.

なお $\rho=0$ の場合, 最尤解の存在定理は既に 仮谷・中村 (1978) [7] に, 最尤解の推定手順およびその実際については, 仮谷 (1975) [4] [5], 仮谷・赤坂 (1977) [6] に述べ

*) *exact* な観測値は, 観測システムの下で観測可能な最小単位を幅にもつ区間データとして取扱う.

られている。

4 Interval-censored data の情報量

Interval-censored data にもとづく，母数の最尤推定量の精度を吟味し，またどのような censoring observation system が望ましいかを考察するために，以下，簡単のため， $\theta=0$ で， T^* が $N(\mu, \sigma^2)$ に従う場合について， $I(T^*; t)$ にもとづく最尤推定量の μ および σ に関する情報量を求めてみよう。

通常の censoring observation system の下では， μ ， σ の最尤推定量 $\hat{\mu}$ ， $\hat{\sigma}$ が存在しない確率は，無視できるほどに小さいが，厳密に言えば 0 ではない。従って $\hat{\mu}$ ， $\hat{\sigma}$ は確率変数ではない。そこで以下の議論は，最尤推定値が存在するという条件の下での，条件つき最尤推定量 $\hat{\mu}$ ， $\hat{\sigma}$ に関するものである。

観測時点列を t_1, \dots, t_k ($t_1 < t_2 < \dots < t_k$) とし， $t = (t_0, t_1, \dots, t_k, t_{k+1})$ とする。ただし T^* は正規分布に従うことを仮定しているので， $t_0 = -\infty$ ， $t_{k+1} = \infty$ である。このとき T^* の関数 $I(T^*; t)$ の確率分布は次のようになる。

$I(T^*; t)$	$(-\infty, t_1]$	$(t_{j-1}, t_j]$	$(t_k, \infty]$	(3)
確率	$\Phi(t_1')$	$\Phi(t_j') - \Phi(t_{j-1}')$	$1 - \Phi(t_k')$	

(11)

ここに $t'_j = (t_j - \mu) / \sigma$. すなわち t'_j は t_j の標準化
度数で, 以後 プライム は常にこの約束に従って用いる.

また $\Phi(\cdot)$ は 標準正規分布の分布関数である.

$$\text{なお, } \delta_j(t^*) = \begin{cases} 1 & t^* \in (t_{j-1}, t_j] \text{ のとき} \\ 0 & \text{その他 のとき} \end{cases} \quad (4)$$

と定義しておけば, censoring observation system $I(T^*; t)$
の下における 確率変数 T^* の尤度関数は 次式によって
与えられる.

$$L(\mu, \sigma; t) = \prod_{j=1}^{k+1} \{ \Phi(t'_j) - \Phi(t'_{j-1}) \}^{\delta_j(t^*)} \quad (5)$$

対数とすれば

$$\log L(\mu, \sigma; t) = \sum_{j=1}^{k+1} \delta_j(t^*) \log \{ \Phi(t'_j) - \Phi(t'_{j-1}) \}. \quad (6)$$

$$\frac{\partial^2 \log L}{\partial \mu^2} = -\frac{1}{\sigma^2} \sum_{j=1}^{k+1} \delta_j(t^*) \left[\frac{t'_j \varphi(t'_j) - t'_{j-1} \varphi(t'_{j-1})}{\Phi(t'_j) - \Phi(t'_{j-1})} + \left\{ \frac{\varphi(t'_j) - \varphi(t'_{j-1})}{\Phi(t'_j) - \Phi(t'_{j-1})} \right\}^2 \right] \quad (7)$$

ここに $\varphi(\cdot)$ は 標準正規分布の密度関数である.

式(7) より μ に関する情報量は

$$i(\mu; t) = -E \left(\frac{\partial^2 \log L}{\partial \mu^2} \right) = \frac{1}{\sigma^2} \sum_{j=1}^{k+1} \left[\frac{t'_j \varphi(t'_j) - t'_{j-1} \varphi(t'_{j-1})}{\Phi(t'_j) - \Phi(t'_{j-1})} + \frac{\{\varphi(t'_j) - \varphi(t'_{j-1})\}^2}{\Phi(t'_j) - \Phi(t'_{j-1})} \right]$$

一方 $t'_0 = -\infty, t'_{k+1} = \infty, t'_0 \varphi(t'_0) = t'_{k+1} \varphi(t'_{k+1}) = 0$ であるから,

$$i(\mu; t) = \frac{1}{\sigma^2} \sum_{j=1}^{k+1} \frac{\{\varphi(t'_j) - \varphi(t'_{j-1})\}^2}{\Phi(t'_j) - \Phi(t'_{j-1})} \quad (8)$$

次に $N(\mu, \sigma^2)$ かゝの大きさ n の無作為標本 (T_1^*, \dots, T_n^*) が、それぞれの観測時点列の下で、 $I(T_1^*; t_1), \dots, I(T_n^*; t_n)$, $t_i = (t_{i0}, t_{i1}, \dots, t_{ik_i}, t_{i k_i + 1})$, $t_{i0} = -\infty$, $t_{i k_i + 1} = \infty$ なる interval-censored data として観測されるとき、この標本のもつ μ に関する情報量は、(8)式を用いて次のように表わすことができる。

$$I(\mu; t_1, \dots, t_n) = \frac{1}{\sigma^2} \sum_{i=1}^n \sum_{j=1}^{k_i+1} \frac{\{g(t'_{ij}) - g(t'_{i,j-1})\}^2}{\Phi(t'_{ij}) - \Phi(t'_{i,j-1})} \quad (9)$$

一方、 $N(\mu, \sigma^2)$ かゝの大きさ n の標本について、exact な値が観測されるとき、この標本のもつ μ に関する情報量は n/σ^2 である。

従つて interval-censored data として観測される標本 (T_1^*, \dots, T_n^*) のもつ μ に関する情報量の効率

$$EI(\mu; t_1, \dots, t_n) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{k_i+1} \frac{\{g(t'_{ij}) - g(t'_{i,j-1})\}^2}{\Phi(t'_{ij}) - \Phi(t'_{i,j-1})} \quad (10)$$

にほぼ等しい。ほぼというのは観測時点列に伴う最尤推定値の存在しない確率が 0 ではないからであるが、普通にはその値は小さく、効率は (10) に等しいとみなすことができる。

$$\begin{aligned} \text{次に } \frac{\partial^2 \log L}{\partial \sigma^2} &= \frac{1}{\sigma^2} \sum_{j=1}^{k+1} \delta_j(t^*) \left[\frac{t'_j(2-t'_j)g(t'_j) - t'_{j-1}(2-t'_{j-1})g(t'_{j-1})}{\Phi(t'_j) - \Phi(t'_{j-1})} \right. \\ &\quad \left. - \left\{ \frac{t'_j g(t'_j) - t'_{j-1} g(t'_{j-1})}{\Phi(t'_j) - \Phi(t'_{j-1})} \right\}^2 \right]. \end{aligned}$$

$$t'_0 g(t'_0) = t'_{k+1} g(t'_{k+1}) = 0, \quad t'^3_0 g(t'_0) = t'^3_{k+1} g(t'_{k+1}) = 0 \quad \text{であるから}$$

上の式より, $I(T^*; t)$ のもつ σ に関する情報量は

$$i(\sigma; t) = -E\left(\frac{\partial^2 \log L}{\partial \sigma^2}\right) = \frac{1}{\sigma^2} \sum_{j=1}^{k+1} \frac{\{t'_j g(t'_j) - t'_{j-1} g(t'_{j-1})\}^2}{\Phi(t'_j) - \Phi(t'_{j-1})} \quad (11)$$

従ってまた, interval-censored data ($I(T_1^*; t_1), \dots, I(T_n^*; t_n)$), $t_i = (t_{i0}, t_{i1}, \dots, t_{ik_i}, t_{ik_i+1})$, $t_{i0} = -\infty$, $t_{ik_i+1} = \infty$ のもつ σ に関する情報量は

$$I(\sigma; t_1, \dots, t_n) = \frac{1}{\sigma^2} \sum_{i=1}^n \sum_{j=1}^{k_i+1} \frac{\{t'_{ij} g(t'_{ij}) - t'_{i,j-1} g(t'_{i,j-1})\}^2}{\Phi(t'_{ij}) - \Phi(t'_{i,j-1})} \quad (12)$$

一方 $N(\mu, \sigma^2)$ からの大きさ n の標本について, exact な値が観測されるとき, この標本がもつ σ に関する情報量は $2n/\sigma^2$ である.

このことから, interval-censored data として観測される標本 (T_1^*, \dots, T_n^*) がもつ σ に関する情報量の効率率は

$$EI(\sigma; t_1, \dots, t_n) = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^{k_i+1} \frac{\{t'_{ij} g(t'_{ij}) - t'_{i,j-1} g(t'_{i,j-1})\}^2}{\Phi(t'_{ij}) - \Phi(t'_{i,j-1})} \quad (13)$$

に等しいとみなすことができる.

[5] Interval-censored data の効率.

exact な値が観測される無作為標本の場合には, 実験または調査すべき標本の大きさ n の見積りについて, 理論的な研究がなされているが, interval-censored data の場合

(14)

には、観測点列のパターンに依存するので簡単ではない。

そこで次には、 $\rho=0$ かつ母集団分布が正規分布 $N(\mu, \sigma^2)$ の場合について、interval-censored data の μ および σ に関する情報量の、exactな標本データに対する効率を算出し、実験または調査計画立案に資することにしよう。なお、以下に示すような基本的な実験・調査計画においては、式(10)、式(13)の値は、 $n=50, 100, \dots, 500$ に対し、ほとんど変わらないことを注意しておこう。

標本個体それぞれに対する観測点列のベクトル t_1, \dots, t_n について次のような規約をもうける。

$t_i = (t_{i0}, t_{i1}, \dots, t_{ik_i}, t_{ik_i+1})$ ($i=1, \dots, n$) において

- a) $k_i = k$ (観測回数はすべてについて k 回),
- b) $t_{ij} - t_{i,j-1} = s$ (>0), ($j=2, \dots, k$) (観測間隔一定),
- c) $t_{ij} - t_{i-1,j} = h$ (>0), ($i=2, \dots, n$) (調査個体の実験への参加は等間隔)。

さらに、 k 回にわたる n 人の全観測点 t_{ij} の中央値を η とし、 $\eta - \mu = d$ (観測点列の母平均 μ からずれる) とする。

$r = (n-1)h$ とおけば、 r は標本個体の年齢幅、または実験への参加時点の広がり を表わしている (図3)。

〈表1〉 $\Sigma_1(50, r, d, 1)$ の下での I-C データの 効率

EFFICIENCY OF INTERVAL-CENSORED DATA 単位 (%)
 ---AMOUNT OF INFORMATION RELATIVE TO μ ---

(AGE RANGE)		AMOUNT OF SHIFT OF MEDIAN(TIJD) RELATIVE TO μ							
$r \backslash d$		0.00	0.50	1.00	1.50	2.00	2.50	3.00	4.00
0.001		63.66	58.10	43.86	26.91	13.11	4.98	1.46	0.06
0.500		63.16	57.71	43.75	27.06	13.37	5.18	1.56	0.06
1.000		61.70	56.57	43.39	27.48	14.12	5.78	1.87	0.09
1.500		59.39	54.76	42.80	28.09	15.27	6.76	2.40	0.15
2.000		56.37	52.39	41.97	28.80	16.73	8.08	3.18	0.25
2.500		52.86	49.60	40.90	29.47	18.34	9.67	4.22	0.42
3.000		49.07	46.53	39.59	30.00	19.97	11.44	5.51	0.69
4.000		41.46	40.17	36.33	30.25	22.76	15.14	8.70	1.66
5.000		34.74	34.23	32.51	29.23	24.33	18.34	12.23	3.41
6.000		29.38	29.22	28.61	27.15	24.43	20.37	15.39	5.95

EFFICIENCY OF INTERVAL-CENSORED DATA 単位 (%)
 ---AMOUNT OF INFORMATION RELATIVE TO σ ---

(AGE RANGE)		AMOUNT OF SHIFT OF MEDIAN(TIJD) RELATIVE TO μ							
$r \backslash d$		0.00	0.50	1.00	1.50	2.00	2.50	3.00	4.00
0.001		0.00	7.26	21.93	30.27	26.22	15.56	6.56	0.45
0.500		0.68	7.61	21.63	29.69	25.93	15.65	6.76	0.50
1.000		2.61	8.59	20.78	28.05	25.06	15.88	7.35	0.65
1.500		5.47	10.06	19.57	25.60	23.67	16.13	8.24	0.92
2.000		8.81	11.81	18.23	22.73	21.84	16.21	9.27	1.35
2.500		12.14	13.61	17.00	19.83	19.72	16.00	10.29	1.96
3.000		15.03	15.25	16.05	17.23	17.49	15.41	11.11	2.77
4.000		18.49	17.43	15.25	13.76	13.49	13.27	11.64	4.80
5.000		18.80	17.81	15.33	12.69	11.13	10.81	10.64	6.79
6.000		17.29	16.78	15.24	12.92	10.63	9.29	9.01	7.83

<表 2> $\Sigma_3(50, r, 10, d, 3)$ の下での, I-C データの効率

SURVEY INTERVAL= 1.000

EFFICIENCY OF INTERVAL-CENSORED DATA

単位 (%)

---AMOUNT OF INFORMATION RELATIVE TO μ ---

(AGE RANGE) AMOUNT OF SHIFT OF MEDIAN(TIJD) RELATIVE TO μ

$r \backslash d$	0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50
0.001	88.24	85.77	77.86	64.06	45.89	27.46	13.23	5.00
0.500	88.03	85.54	77.60	63.86	45.88	27.65	13.50	5.21
1.000	87.39	84.83	76.83	63.29	45.84	28.21	14.29	5.82
1.500	86.30	83.65	75.59	62.39	45.80	29.06	15.53	6.82
2.000	84.72	82.00	73.95	61.26	45.74	30.14	17.12	8.17
2.500	82.64	79.89	71.98	59.98	45.70	31.34	18.94	9.83
3.000	80.05	77.36	69.77	58.62	45.65	32.57	20.89	11.71
4.000	73.52	71.22	64.91	55.88	45.52	34.87	24.73	15.85
5.000	65.78	64.19	59.73	53.12	45.20	36.68	28.09	19.94
6.000	57.87	57.01	54.42	50.15	44.46	37.82	30.71	23.55

EFFICIENCY OF INTERVAL-CENSORED DATA

単位 (%)

---AMOUNT OF INFORMATION RELATIVE TO σ ---

(AGE RANGE) AMOUNT OF SHIFT OF MEDIAN(TIJD) RELATIVE TO μ

$r \backslash d$	0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50
0.001	54.06	49.52	40.81	36.45	37.58	36.61	28.17	16.01
0.500	53.63	49.33	41.02	36.72	37.48	36.25	28.00	16.15
1.000	52.42	48.77	41.61	37.46	37.20	35.25	27.51	16.53
1.500	50.66	47.94	42.40	38.47	36.79	33.83	26.76	17.03
2.000	48.64	46.95	43.17	39.47	36.32	32.25	25.83	17.53
2.500	46.69	45.90	43.67	40.22	35.85	30.79	24.85	17.91
3.000	45.04	44.89	43.74	40.52	35.39	29.62	23.97	18.13
4.000	43.00	43.08	42.41	39.56	34.42	28.39	22.90	18.21
5.000	41.90	41.37	39.80	37.09	33.10	28.09	22.93	18.42
6.000	40.25	39.33	37.07	34.33	31.34	27.75	23.52	19.19

〈表 3〉 $\Sigma_3(50, r, 30, d, 3)$ の下での I-C データの効率

SURVEY INTERVAL = 3.000

EFFICIENCY OF INTERVAL-CENSORED DATA

単位(%)

---AMOUNT OF INFORMATION RELATIVE TO μ ---

(AGE RANGE)		AMOUNT OF SHIFT OF MEDIAN(TIJD) RELATIVE TO μ							
$r \backslash d$		0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50
0.001		65.33	61.61	54.06	50.22	54.03	61.45	64.50	58.26
0.500		64.99	61.43	54.23	50.58	54.20	61.25	64.07	57.89
1.000		64.00	60.92	54.72	51.60	54.68	60.67	62.85	56.82
1.500		62.53	60.17	55.46	53.09	55.39	59.79	60.96	55.14
2.000		60.81	59.30	56.32	54.82	56.19	58.69	58.59	53.00
2.500		59.10	58.45	57.17	56.49	56.94	57.48	55.98	50.57
3.000		57.66	57.74	57.88	57.88	57.47	56.23	53.37	48.05
4.000		56.21	57.00	58.55	59.09	57.46	53.73	48.83	43.45
5.000		56.66	57.18	58.12	58.01	55.71	51.26	45.71	40.22
6.000		57.87	57.66	56.93	55.39	52.68	48.66	43.67	38.46

EFFICIENCY OF INTERVAL-CENSORED DATA

単位(%)

---AMOUNT OF INFORMATION RELATIVE TO σ ---

(AGE RANGE)		AMOUNT OF SHIFT OF MEDIAN(TIJD) RELATIVE TO μ							
$r \backslash d$		0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50
0.001		13.13	25.85	51.93	65.26	51.56	24.01	6.57	9.10
0.500		14.32	26.48	51.34	63.99	50.93	24.50	7.50	9.58
1.000		17.69	28.26	49.65	60.40	49.10	25.86	10.15	10.99
1.500		22.72	30.87	47.12	55.10	46.30	27.75	14.10	13.18
2.000		28.65	33.87	44.13	48.96	42.90	29.76	18.73	15.93
2.500		34.54	36.79	41.13	42.88	39.31	31.45	23.34	18.98
3.000		39.52	39.23	38.54	37.61	35.93	32.52	27.28	21.99
4.000		44.50	41.60	35.55	31.34	31.09	32.48	31.52	26.70
5.000		42.67	40.41	35.28	30.86	29.55	30.51	30.87	28.29
6.000		37.54	37.14	35.72	33.23	30.49	28.67	27.89	26.85

観測システム Σ_3 になると, σ が大きく影響して複雑になり, とても直観的に妥当な観測システムを策定することは困難であることがわかるであろう。 μ の推定に重点をおくか, σ の推定に重点をおくか, あるいは両者の推定値に ほぼ同等の精度を望むかによって 観測計画は大幅に違ってくるので, 表を 入念に検討して 計画を策定することが必要である。

1974年4月下旬, A 中学校新入生 約 500 名の女生徒を対象とした 初潮調査 Σ_1 (500人, 365日, 40日, 1) で μ, σ に関する情報量の効率 は それぞれ 62%, 2.1% であった。この調査は μ の推定に重点をおいた調査であったので, ほぼ満足のいく精度であったということができる。 $500 \times 0.62 = 310$ であるから, 「初潮は経験済みか否か」だけを 1 回に限って調査したこの調査の, μ に関する情報量としては, 約 310 人についての正確な初潮年令(単位日)データに匹敵していることがわかる。もし σ に重点がおかれていたとすれば, $500 \times 0.021 = 10.5$ 人の正確なデータに匹敵する情報量しか含んでいないので, もう少し時期をずらすことが必要であることが <表1> から読みとれる。

参 考 文 献

- [1] Boag, J. W. (1949); Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy, Journal of the Royal Statistical Society, Vol. B11, p.15-53.
- [2] Harris, T. E., P. Meier, and J. W. Tukey (1950); The Timing of the Distribution of Events Between Observations, Human Biology, Vol. 22, p. 249-270.
- [3] Hughes, E. J. (1962); Maximum Likelihood Estimation of Distribution Parameter from Incomplete Data, Universty Microfilms Inc., Ann Arbor.
- [4] Kariya, T. (1975); Menarche Age Distribution Estimated by the Method of Maximum Likelihood on Incomplete Quantal Response Data, Kawasaki Medical J., Vol. 1, p.75-83. (in Japanese)
- [5] Kariya, T. (1975); Incomplete Quantal Response Data Analysis-Maximum Likelihood Estimation of Parameters Based on Mixture of Interval Data and Ordinary Ones, Kawasaki Med. J., Vol. 1, p.85-93.
- [6] Kariya, T. and M. Akasaka (1977); The Eruption Age Distribution of Deciduous Teeth-Maximum Likelihood Estimation of Parameters Based on Incomplete Quantal Response Data, Japanese J. Applied Statistics, Vol. 5, p.3-18. (in Japanese)
- [7] Kariya, T. and T. Nakamura (1978): The Maximum likelihood Estimates Based on the Incomplete Quantal Response Data, Journ. Japan Statist. Soc., Vol. 8, 1, p.21-28.
- [8] Kulldorff, G. (1958a); Maximum Likelihood Estimation of the Mean of a Normal Random Variable when the Sample is Grouped, Skand. Aktuarietidskr., Vol. 41, p.1-17.

[9] Kulldorff, G. (1958b); Maximum Likelihood Estimation of the Standard Deviation of a Normal Random Variable when the Sample is Grouped, Skand. Aktuarietidskr., Vol. 41, p.18-36.

[10] Peto, R. (1973); Experimental Survival Curve for Interval-Censored Data, Appl. Statist., Vol. 22, p. 86-91.